Fixation Prediction and Visual Priority Maps for Biped Locomotion

N. Anantrasirichai, *Member, IEEE,*, K. A. J. Daniels, J. F. Burn, Iain D. Gilchrist, and David Bull, *Fellow, IEEE*

***Abstract*—This paper presents an analysis of the low-level features and key spatial points used by humans during locomotion over diverse types of terrain. Although, a number of methods for creating saliency maps and task-dependent approaches have been proposed to estimate the areas of an image that attract human attention, none of these can straightforwardly be applied to sequences captured during locomotion, which contain dynamic content derived from a moving viewpoint. We used a novel learning-based method for creating a visual priority map informed by human eye tracking data. Our proposed priority map is created based on two fixation types: firstly exploiting the observation that humans search for safe foot placement and secondly that they observe the edges of a path as a guide to safe traversal of the terrain. Texture features and the difference between them, observed at the region around an eye position, are employed within a support vector machine to create a visual priority map for biped locomotion. The results show that our proposed method outperforms the state-of-the-art, particularly for more complex terrains, where achieving smooth locomotion needs more attention on the traversing path.**

***Index Terms*—Salience, priority map, locomotion, eye tracking, bio-inspired.**

I. Introduction

VISION provides us with information that can be used for adaptively controlling our locomotion. However, we still do not fully understand how humans perceive and use this information in a dynamic environment. Eye tracking can be used to capture the deployment of the high resolution fovea on an instant-by-instant basis which is key in understanding what visual information is important in different visual contexts [1], [2]. Most previous research has studied eye movements in the context of visual search tasks using static images with participants being asked to look at a series of images on a computer screen. Other research has employed portable/wearable eye trackers to acquire videos of gaze fixations while the participants were walking [3], [4]. However, the main focus of these investigations has been on the detection of new events occurring during locomotion, e.g. when humans approach obstacles [5], [6], encounter different ground terrain [7], changed direction [8], or encounter other moving objects while walking [9]. None of these studies have focused on fixations related to continuous terrains during locomotion. Combining these approaches could offer a complete bio-inspired solution for autonomous robots to make locomotion decisions when traversing difficult and varied terrain. It has potential application in the context of guidance aids for the visually impaired.

In this paper, visual information provided by eye movements, captured from the viewpoint of human locomotion, is studied during walking and running over a range of different types of terrain. The human visual system provides a sense of distance, global information about self-motion through an environment and the posture of the body relative to the environment [10]. Understanding the features and key points exploited by humans could therefore improve the performance of autonomous systems, where cameras are frequently employed as primary sensors, to emulate the way human eyes perceive the navigable environment. Here we model a *priority map* in order to predict eye positions. The priority map reflects the combined representations of salience (bottom-up) and relevance (top-down) in the selection process, which best describe the firing properties of neurons in the visual cortex [11], [12]. Salience is the property of a scene, where specific features combine to attract visual attention, while relevance exploits top-down factors, e.g. expectation and experience, to determine attentional allocation. In this paper, which focuses on maintaining smooth locomotion under varying terrain conditions, both bottom-up and top-down processes are employed to recognize objects, material types and surface conditions, associated with visually guided behaviour. To create a priority map, we employ machine learning using a support vector machine (SVM) [13] for probability estimation of gaze location. Key points on the map with high probabilities can then be used to control movement and path planning.

The remaining part of this paper is organized as follows. Section II presents existing work on constructing a visual priority map. Eye-tracked sequences on human locomotion are then discussed in Section III. Subsequently, we describe our proposed method of saliency estimation in Section IV and the results are shown in Section V. Finally the conclusions and future work are set out in Section VI.

II. Existing visual selection models

As we perceive a great amount of visual information constantly, our nervous system makes decisions on which parts of this information should be further processed, and

This work has been supported by Engineering and Physical Sciences Research Council (EPSRC) grant EP/J012025/1 and EP/M000885/1.

N. Anantrasirichai and David Bull are with Visual Information Laboratory, University of Bristol, Bristol BS8 1UB, U.K (e-mail: N.Anantrasirichai@bris.ac.uk; Dave.Bull@bris.ac.uk).

K. A. J. Daniels and J. F. Burn are with Department of Mechanical Engineering, University of Bristol, Bristol, UK (e-mail:J.F.burn@bristol.ac.uk; K.Daniels@bristol.ac.uk).

Iain D. Gilchrist is with School of Experimental Psychology, University of Bristol, Bristol, UK (e-mail: I.D.Gilchrist@bristol.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier

also prioritizes it. Saliency-based modeling, taking account of human visual attention and visual search strategies, is one of the successful approaches to this problem. Beginning with the influential bottom-up method by [14] that replicated the early process in the primate visual system, this was followed by the work of Koch and Itti and their research teams who proposed a number of successively more refined saliency models [15]–[18]. More recently task-dependent approaches (top-down) have been included to deal with complex environments [19].

Itti et. al. [15] modelled visual attention based on the feature integration theory, where elementary features, e.g. colour, intensity and orientations represented in the visual cortex, were processed in parallel and in a multi-scale manner to distinguish the objects presented. They also included centre-surround processes, inspired by neural responses in the lateral geniculate nucleus (LGN) [14], in which image values in a centre-location to its neighbouring surround-locations are compared. Hou et.al. [18] introduced a saliency map based on image signature that spatially approximated the foreground of an image and predicted fixation points using a discrete cosine transform (DCT). Multiple scales were employed through a set of weighted centre-surround outputs in [20]. Zhang et al. modelled a Bayesian framework from the self-information of visual features, and overall saliency emerged as the pointwise mutual information [21]. A salient object was detected using a wavelet transform associated to human visual system in [22], whilst a phase filter computed using a discrete Fourier transform (DFT) was employed in [23].

Instead of using a set of biologically plausible filters, some models have been created by training a classifier directly from human eye tracking data. In [24], saliency was determined by quantifying the joint likelihood and self-information of each location image patch. A large samples of random patches of natural images were employed in an independent component analysis process to find a suitable basis. In [25] and [26], three levels of features were employed, namely low-level features, mid-level features such as the objects at the horizon, and high-level features such as people. Liang et. al. [27] selected key low-level features using the SVM classifier to create a saliency map based on eye fixations. Principal component analysis (PCA) was employed to separate targets from peripheral regions in [28]. Convolutional neural networks (CNNs) have been widely employed to process visual and other two-dimensional data. With a large amount of available eye movement data, CNN-based methods have been used for fixation prediction, creating a saliency map via the softmax function [29], [30].

Three dimensional (3D) data has also been employed, based on the fact that humans perceive visual information using both the current scene (spatial information) and accumulated knowledge (temporal information). Motion was captured using directional masks [31], optical flow [32], 3D textures [33]. All of these methods aim to detect moving objects against static or slow panning background. Recurrent neural networks (RNNs) have been employed to perform sequence recognition by providing at least one feed-back connection. The most commonly used type of RNN is the long short-term memory (LSTM) network, as this solves the vanishing gradient problem, observed in traditional RNNs by memorising sufficient context information in time series data via its memory cell [34]. An LSTM network has been combined with a CNN to detect saliency in 3D data in [35], [36]. 2D convolution operators were applied to extract key features or spatial salience, and then fed to the LSTM network to capture the sequence of actions.

An intensive review of visual attention modelling can be found in [37]. These methods were formulated from a benchmark data set of eye-movement fixation points. However, they were not developed for vision information received during locomotion, and hence unfortunately do not fully apply to this scenario. Note that salience is a relative property referring in a stimulus-driven process without influenced by cognitive top-down factors, e.g. expectation and experience, or goals of the observers. However, the terms 'salience' and 'relevance' are sometimes interchangeable in the neurophysiological literature. Some research constructs an integrated framework for attentional selection and called it 'priority' map as the stimulus-based salience map interacts with the cognitive factors to guide visuomotor behaviour [11], [38]. We hence refer our results as priority maps, where the areas and key points on the image are prioritized based on local features and the goal of smooth locomotion.

III. Analysis of terrain sequences with an eye tracker

The test sequences used in our work were acquired with a mobile eye tracker that produces a point of view video at a resolution of 1280×960 pixels ($W \times H$) at 24 fps, as well as a record of eye position recorded at 30 fps. The system typically delivers a gaze tracking range of 80° horizontally and 60° vertically, while providing a scene field of view of 60° horizontally and 46° vertically. The average error of the eye tracker measured in the field using recalibration is 1.68 degree with the standard deviation of 0.76. Eight participants were asked to walk on *flat concrete*, *slant cobbles*, *rocks* and *stepping stones* (regular protrusions of the *flat concrete* path surface into the rock region spaced approximately one step length apart) alongside the Severn Way footpath at Severn Beach, South Gloucestershire, UK [1]. Data were also collected while running on *flat concrete*, *slant cobbles* and *stepping stones* but not on the *rocks* as the risk of falling was perceived to be too high. Six participants had never visited the location before, while two participants had visited previously but did not spend time there regularly. Fig. 1 shows some examples of the scenes.

Some key observations can be made as follows.

1) Relatively few eye movements were directed to the path of travel when walking over *flat concrete*. Participants generally looked at other objects, scenery, and pedestrians, or to a more complex area of terrain they were aware they were about to traverse. Eye movements were variable and often large. This behaviour also occurred occasionally when walking on *slant cobbles*.

[1]Dataset can be accessed at eis.bris.ac.uk/∼eexna/download.html

Fig. 1. Severn Beach scenes with 4 terrain difficulties. Top-left: flat concrete. Top-right: slant cobbles. Bottom-left: stepping stones. Bottom-right: rocks.

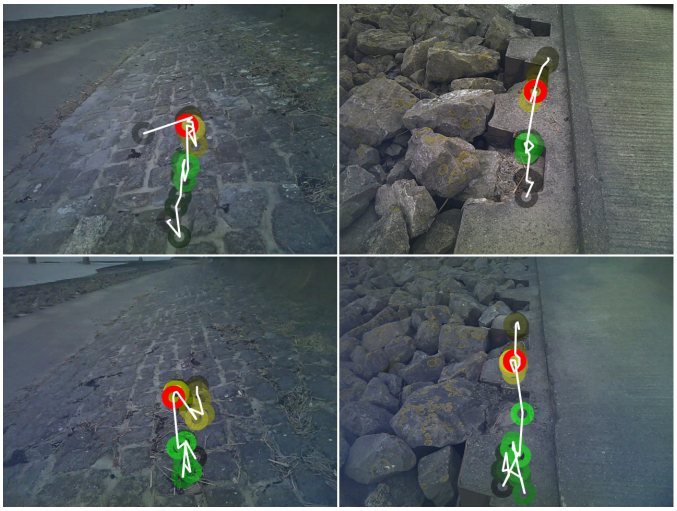

Fig. 2. Examples of 'jump up and then track back behaviour'. Red is the current eye position. Greens and yellows represent the previous and future points in times, respectively. White lines connect the eye positions between successive frames.

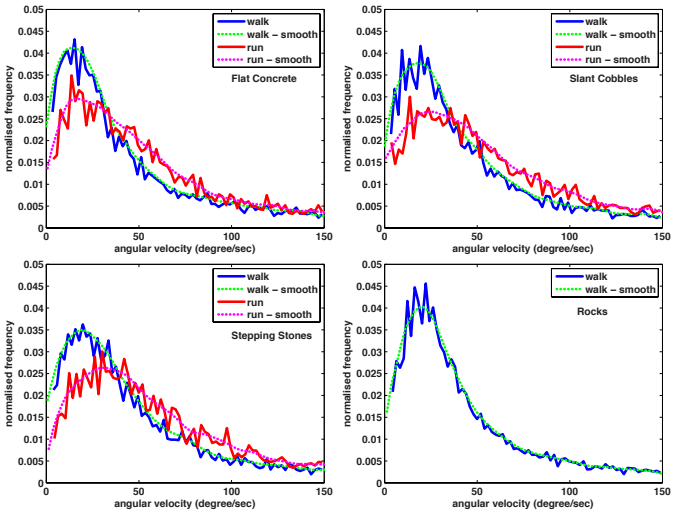

Fig. 3. Normalized histograms of angular velocity (°/s) of eye movements while walking and running. The angular velocity was computed from two points, for which the point of the previous frame was warped to the current frame's geometry to compensate head movement. The data of running on the rocks was not collected according to high risk.

2) On *slant cobbles* and *stepping stones*, eye movements were predominantly vertically oriented. Smooth downward movements, apparently maintaining fixation on an environmental feature, were separated by fast upward movements ('*track and return*' behaviour) [39]. Examples of this behaviour are shown in Fig. 2, where eye positions over one second (15 previous frames (green) and 15 future frames (yellow)) are registered to the current frame (red). This shows that our eyes fixate a particular location tracking it back as walking forwards, then saccading ahead again to fixate the next location.
3) When running, the same '*track and return*' behaviour was apparent on *slant cobbles* and *stepping stones* as observed while walking. This behaviour also appeared, at least some of the time, when running on *flat concrete*.
4) Fig. 3 demonstrates the frequencies of different angular velocities of the eye movements for each terrain while walking and running. An angular speed less than $2^\circ/s$ (where the dip in the plot occurs if the whole data is employed to generate the histogram) is defined as a fixation and not included in the plot. The most frequent angular velocity is approximately $20^\circ/s$ and $30^\circ/s$ while walking and running, respectively. The peak angular speed can be as fast as $400^\circ/s$ and $500^\circ/s$ again while walking and running, respectively. The eyes move more often and faster while running, and also scan larger areas of the scene. This is probably because shorter period is available to make decisions when moving faster, so more information of the scene is assembled for planning.
5) Eye movements on the *stepping stones* were very consistent. Fixations were oriented approximately two steps ahead and fast eye movements were used to increment gaze position between consecutive stepping stones during locomotion.
6) On the *stepping stones*, fixations were often at the boundaries of the stepping path rather than oriented to safe areas for foot placement. This may be because neurons in visual areas are sensitive to texture boundaries and higher contrast edges [40]. Marigold and Patla suggested that fixations on the transition region are more related to gathering greater amounts of information about the terrain characteristics and layout rather than for guiding precise foot placement [7].
7) On *rocks*, more lateral eye movements were observed. This appeared to be part of a search for safe foot placements by fixating both edges of the *rocks* and flatter areas. Participants often appeared to only be looking one step ahead and occasionally paused locomotion while searching for the next foot placement. Re-fixations were sometimes observed and areas that were used for foot placement had almost always previously been fixated. Eye movements were generally made systematically to the

future location of the next foot placement but sometimes other possible foot placement locations that did not get used for support were fixated.

8) It appeared that head pitch angle was more downward when walking over more complex and difficult terrain (also previously noted for eye-in-head and head-in-world orientation [39]). Eye movements were centralized as shown in Fig 4, in agreement with the findings of [41].

From observations 1, 5, 7 and 8 it can be concluded that eye fixation patterns are highly dependent on terrain difficulty, i.e. they are task-relevant. Observations 6 and 7 reveal that there are two fixation types, indicating that humans search for locations to ensure safe foot placement (Fixation type 1) and observe the edges of the path to guide their path through safe terrain (Fixation type 2). Combining these two groups of observations, we can create two saliency maps, for fixation types 1 and 2 separately, and merge them with a weight relative to terrain difficulty, based on the means and variances of the high frequency image content, to create a final priority map (see Eq. 9 in Section IV-B). Observations 1 and 4 suggest that the oncoming path is frequently checked - possibly for planning, while the immediate area may be repeatedly viewed - possibly to ensure safety, particularly on challenging terrains or at increased speed. Observation 7 on *rocks* is also demonstrative of fixations for path planning (although this is very local path planning). Observation 8 reveals that the eye positions obtained from the videos of this experiment exhibit centre-bias behaviour - since the head is often moved to improve vision. However, in many practical robotic systems, camera angle is typically fixed. Therefore, we do not give automatically more weight to the centre of the image in our approach. Besides this, a centre-bias assumption may limit the area of saliency point detection in more distant regions, where the participants have been shown to fixate occasionally - as demonstrated by observations 1, 2 and 3.

IV. Proposed Visual Priority Map for Biped Locomotion

The observations from the previous section are employed to form a framework to model fixation behaviour. The proposed scheme is shown in Fig. 5. The blue part of the figure represents the learning process, which comprises two models – one for fixation at safe areas of foot placement and the other for fixation at the edges of objects or terrain. The pink part of the figure, where some sub-processes overlap the learning process, shows the process for creating a priority map for a current frame. Details of each step are described below.

A. Training process

First, we discard saccades using the approach in [42] and the blurred frames as follows. When the shutter speed of a camera is not fast enough to capture stop motion, some frames will exhibit high levels of motion blur which may alter measured frequency properties. The sharpness value of each frame is computed from the mean of high-pass magnitudes as follows.

$$\Psi = \frac{1}{n_{all}}\left(\sum_{l=1}^{4}\sum_{s=1}^{6}\sum_{i=1}^{n_{s,l}}|\psi_{l,s,i}|\right)/n_{all} \qquad (1)$$

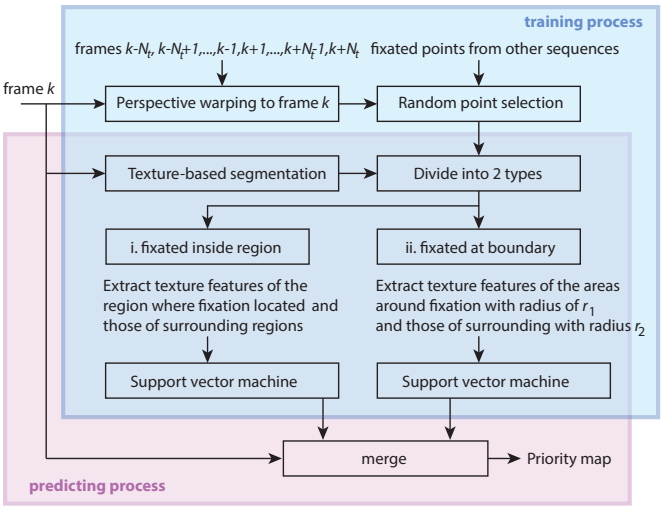

Fig. 5. Diagram of proposed priority map algorithm for human locomotion. Blue box shows the training process, while the pink box demonstrates the estimation process

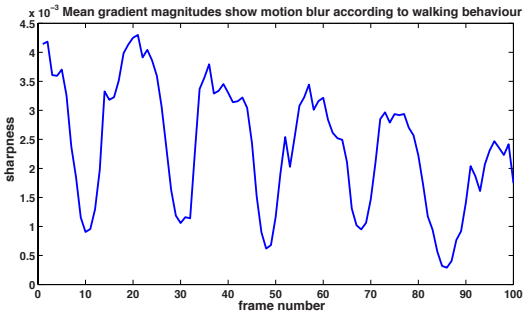

Fig. 6. Sharpness shows walking step

where $\psi_{l,s,i}$ is a wavelet coefficient i of subband s at decomposition level l. $\psi_{l,s,i}$ is obtained from the Undecimated Dual-Tree Complex Wavelet Transform (UDT-CWT) [43]. $n_{s,l}$ and n_{all} are the number of wavelet coefficients of each subband and the total number of all levels, respectively. Fig. 6 clearly shows the points where the camera is moving faster. This indicates when the body vaults over the leg at each step during normal walking. Frames are retained if their sharpness values are higher than a threshold τ_{sharp}, which is adaptively defined using N_b backward and N_f forward frames. $\tau_{sharp,k}$ for frame k is computed as in Eq. 2.

$$\tau_{sharp,k} = \frac{\alpha}{N_b + N_f + 1}\sum_{n=k-N_b}^{k+N_f}\Psi(n) \qquad (2)$$

Here α is set to 0.8. If $\Psi_k > \tau_{sharp,k}$, frame k is used.

1) Texture-based segmentation: Each frame, where $\Psi_k > \tau_{sharp,k}$, is segmented into non-overlapping regions for which the texture characteristics of adjacent areas are different. We employ a wavelet-based watershed segmentation [44], but its gradient map is generated using the UDT-CWT, in a similar manner to the sharpness calculation and the texture features used in probability estimation process. The output is a segmentation map Ω_k. To reduce computational time, we resize the image by the factor of 0.25 before applying the

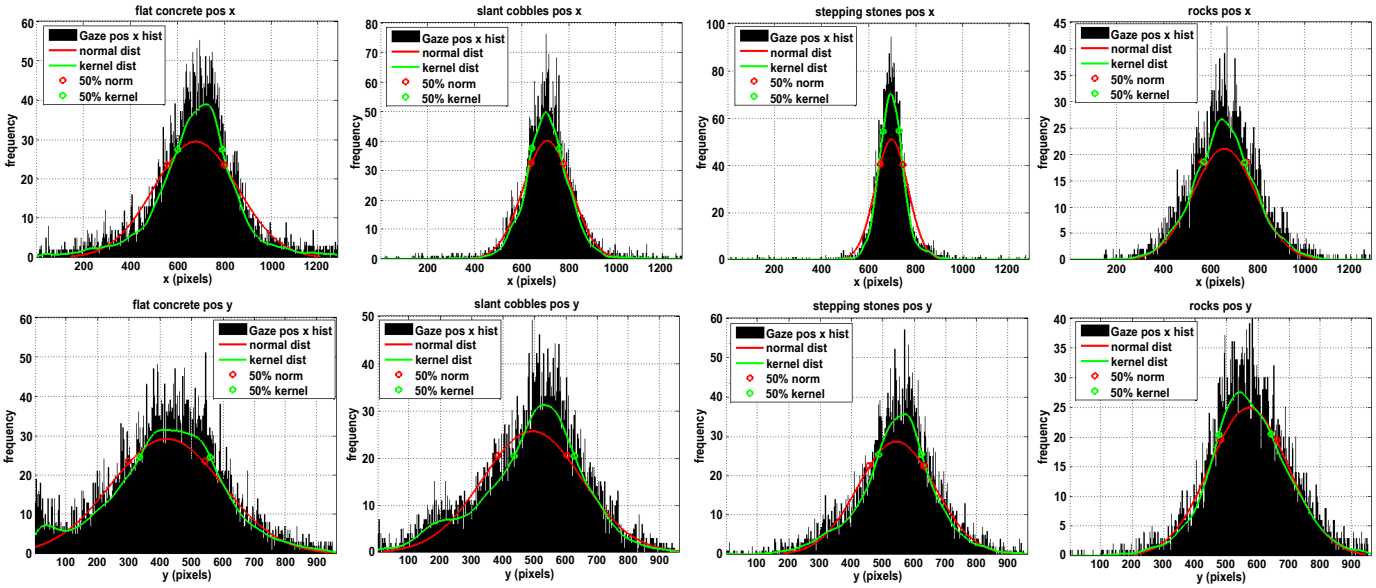

Fig. 4. Distributions of eye positions on the horizontal (Top) and the vertical (Bottom) directions of (Left to Right) flat concrete, slant cobbles, stepping stones and rocks. The coordinate (0,0) is at top left of the image.

segmentation process. The Ω_k is subsequently interpolated to the original image size.

2) Fixation types: Eye tracking sequences reveal that human fixations typically occur in two areas on an image. The first type is well-established in the literature indicating that humans look where they are going to step [2], [3]. The second type relates to fixations at terrain boundaries (as clearly noticeable in the *stepping stones* and *rocks* sequences) and is normally associated with more complex environments. Data from each case is employed to train the SVM classifier separately.

We employ Ω_k to classify the eye positions to either the first type $c_k = 1$ or the second type $c_k = 2$ as described in Eq. 3.

$$c_k = \begin{cases} 1, & \text{if } d_{L_k} > \tau_{L_k} \\ 2, & \text{if } d_{L_k} \leq \tau_{L_k} \end{cases} \tag{3}$$

where d_{L_k} represents a distance from the eye position p_k to the nearest point on L_k, and τ_{L_k} is a defined threshold. The boundary lines L_k between regions in Ω_k are the edges of the objects, or where different textural terrains meet. Therefore, Ω_k can be employed to define fixation type. If p_k is inside the region – it is not too near to the boundary line, p_k would be of the first type ($c_k = 1$). If p_k is located near L_k, it implies that the participant is aware of the unsafe place ($c_k = 2$). This idea is illustrated in Fig. 7.

3) Outliers: Sometimes a participant will look somewhere not related to the path of locomotion, particularly when walking on simple terrain. These fixations are ignored here, because their primary purpose is not to assist with locomotion. We remove outliers using a histogram approach. Fixation locations that are distant from the majority are removed. First, histograms of x and y components of the eye positions are created. Instead of using a normal distribution to estimate the majority locations, we employ kernel density estimation, which is a non-parametric way to estimate the probability

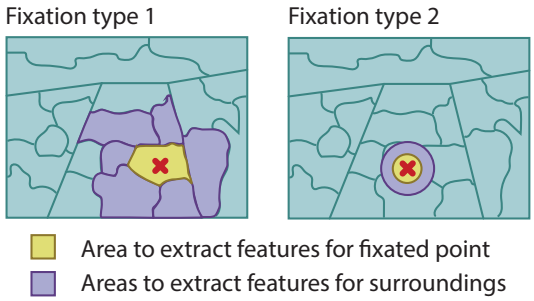

Fig. 7. Region-based feature extraction for each fixation type

density function of a random variable as shown in Eq. 4.

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{4}$$

where n is the sample size, $K(\bullet)$ is the kernel smoothing function, $h > 0$ is the bandwidth (smoothing parameter). If p_k is not within 80% of the majority of the population, it is removed. This percentage has to be small enough to remove eye positions that do not fixate on the ground. This was empirically determined using the *flat concrete* sequences. Fig. 4 displays the histograms of eye positions for each terrain type. It reveals that participants look up sometimes while walking on simple terrains as the histograms of the vertical directions of *flat concrete* and *slant cobbles* show a second peak near the top part of the image. Moreover, the frames that show saccadic eye movement larger than 20°/s and 30°/s are removed from walking and running sequences, respectively, following the Observation 4 in Section III.

4) Feature extraction: The texture of the area around the eye fixation position p_k is extracted. Texture is an efficient tool for characterizing various material properties, such as structure, orientation, roughness, smoothness, and regularity

TABLE I
LIST OF TEXTURE FEATURES (υ_r^k)

Features	# dimensions
Intensity level distribution (ILD)	
Mean, Variance, Skewness,	
Kurtosis, Entropy	5
Complex wavelet transform (CWT)	
(4 decomposition levels)	
Mean, Variance of magnitudes	8
Mean of magnitudes of each sub-band	48
Local Binary Pattern (LBP)	
Histogram	59

differences within an image. Texture features used for this paper are given in Table I. This set of features were found to be the best combination for terrain classification [45]. Only the intensity (Y) channel, extracted from the $\mathrm{YC}_b\mathrm{C}_r$ colour transformation, is used here.

For the *intensity level distribution*, five parameters are extracted, including mean, variance, skewness, kurtosis and entropy. The *Local binary pattern* labels the pixels in an image by thresholding the neighbourhood of each pixel, considering the result as a binary number [46]. Uniform patterns are generated using 8 sampling points on a circle of radius 1 pixel. There are a total of 256 patterns, 58 of which are uniform, which produce 59 output labels. A histogram with 59 bins is obtained, and the frequency of each bin is used as one feature.

As one of the most important aspects of texture is scale, which provides both spatial and frequency information, a multi-resolution approach is utilized based on *wavelet features*. Wavelets have been extensively used to extract spatial frequency (e.g. edges and lines) and spatial orientation, since their mathematical properties fit well with those of the early visual system, e.g. two-dimensional receptive field profiles are well described by two-dimensional Gabor functions [47]. We employ the Undecimated Dual-Tree Complex Wavelet Transform (UDT-CWT) [43] which uses two different real discrete wavelet transforms (DWT) to provide the real and imaginary parts of the CWT without sub-sampling. This increases directional selectivity over the DWT and is able to distinguish between positive and negative orientations giving six distinct sub-bands at each level, corresponding to $\pm 15°$, $\pm 45°$, $\pm 75°$. This also provides shift-invariance and good directional selectivity. The undecimated version is employed because it does not suffer from the problem of mis-alignment between features in the image and features in the bases, resulting better priority map construction. With 4 decomposition levels, the mean and variance of magnitudes across all subbands in each region produced 8 features and those of each subband produce further 48 features (2×4 levels$\times6$ subband/level).

The final list of features employed to predict the eye position are denoted f_k and g_k. f_k is a list of texture features and g_k is the difference between texture features of the eye position and those of its neighbouring areas, following the observation that fixations occur where there are large differences in surrounding textures [14], [48].

The areas used in feature extraction for each fixation type are shown in Fig. 7. For $c_k = 1$ where eye position is inside the region r, texture features of r are extracted, denoted $\boldsymbol{\Upsilon}_r^k = \{\upsilon_{r,1}^k, \upsilon_{r,2}^k, ..., \upsilon_{r,N_f}^k\}$, where the N_f represents the total number of texture features (120 in this paper). If there are N_s regions around r, the texture features of these neighbouring regions are $\boldsymbol{\Upsilon}_{r,s}^k, s \in [1, N_s]$. For $c_k = 2$ where eye position is on or near L_k, indicating fixation near terrain boundary, the traditional centre-surround scheme is applied. Texture features are extracted from the area within the radius of ρ_1 from p_k and the area of the ring with radius between ρ_1 and ρ_2, denoted $\boldsymbol{\Upsilon}_{\rho_1}^k$ and $\boldsymbol{\Upsilon}_{\rho_2}^k$ for centre and surrounding areas, respectively. That is, the features f_k and g_k for this eye position p_k are described as in Eq. 5 and Eq. 6.

$$f_k = \begin{cases} \boldsymbol{\Upsilon}_r^k, & \text{if } c_k = 1 \\ \boldsymbol{\Upsilon}_{\rho_1}^k, & \text{if } c_k = 2 \end{cases} \tag{5}$$

$$g_k = \begin{cases} \left| \boldsymbol{\Upsilon}_r^k - \frac{1}{N_s} \Sigma_{s=1}^{N_s} \boldsymbol{\Upsilon}_{r,s}^k \right|, & \text{if } c_k = 1 \\ \left| \boldsymbol{\Upsilon}_{\rho_1}^k - \boldsymbol{\Upsilon}_{\rho_2}^k \right|, & \text{if } c_k = 2 \end{cases} \tag{6}$$

Raw features are normalized so that large values do not dominate in the classification processes. We denote the γ-th feature as $\mathbf{g}_\gamma = \{(f_{\gamma,1}\ g_{\gamma,1}), (f_{\gamma,2}\ g_{\gamma,2}), ..., (f_{\gamma,N}\ g_{\gamma,N})\}$, where $\gamma \in \{1, ..., N\}$ and N is the total number of samples. Note that N may be less than the total number of frames on the sequence because blurred frames are unused. The normalized version $\mathbf{g''}_\gamma$ is computed using max-min scaling to a range of $0-1$, and then they are non-linearly adjusted to give priority to the feature values around the mean using a sigmoid function as described in Eq.7. With this method, feature values are more evenly distributed between 0 and 1 (not concentrated near 0 or 1 because of outliers). Our experimental results show that including this non-linear adjustment increases classifier performance accuracy by up to 2%.

$$\mathbf{g'}_\gamma = (\mathbf{g}_\gamma - g_{\gamma,min})/(g_{\gamma,max} - g_{\gamma,min}) \tag{7a}$$

$$\mathbf{g''}_\gamma = \mathbf{g'}_\gamma^2/(\bar{\mathbf{g'}}_\gamma^2 + \mathbf{g'}_\gamma^2) \tag{7b}$$

5) Perspective warping and random point selection: In order to apply a support vector machine, data must be divided into two classes – here chosen to be fixations and random points. Following the recommendation in [49], the location of the negative sample for the current frame k is randomly selected from one of the fixation locations of all training sequences with the same locomotion type, e.g. walking/running. This is to ensure that fixated and non-fixated distributions have the same bias and the differences between them are not simply selected because participants tend to fixate more to the centre of images. Then, features of the random point of the current frame k are extracted from the area at this location.

We select a random point q_k for the current frame k, using the eye position from a different eye-tracked sequence over the same terrain type, denoted sequence Q. q_k must be a specified distance from the eye positions in frames $k-N_b$ to $k+N_f$ of the current sequence. To achieve this, the eye positions p_m in the neighbouring frames, where $m \in [k-N_b, k+N_f], m \neq 0$, are warped to the current frame k using optical flow estimated using the RANSAC technique [50]. $p_{k,m}$ is the eye position of frame m in the geometry of the current frame k, i.e. $p_{k,m} = w_{m\to k}(p_m)$, where $w_{a\to b}(x)$ is a warping function from frame a to frame b and x is the location in frame a. If q_k is within a

radius ρ_1 of any $p_{k,m}$, it will be discarded. The eye positions of the nearest neighbouring frames in the sequence Q are then checked until one of them maps to a position which has a separation greater than ρ_1. This position will be used as q_k.

6) Probability estimation by support vector machine: We employ a support vector machine (LIBSVM) [13] to perform linear classification and compute the probability. The linear kernel is robust to overfitting and gives better speed than a non-linear kernel. We train the classifiers by labelling the instant eye position in the current frame as positive, and labelling the random points as negative.

B. Priority map construction

A priority map S_k is constructed using the models generated following Section IV-A. The process is illustrated in the pink section of Fig. 5. In real-time applications, the forward frames do not exist, i.e. $n \in [k - N_b, k]$. Hence a more intelligent technique is required to predict the sharp and blurred frames in a walking cycle without knowing the future frames, e.g. our technique proposed in Section V of [45]. This technique also ensures that the change of terrain characteristics from high detail texture, such as grass and bricks, to low detail texture, such as smooth tarmac, will not cause over skipping.

Each selected frame is segmented using texture-based segmentation, producing N_r regions. The texture features, $f_r^{c=1}$ and $g_r^{c=1}$, of each region $r \in [1, N_r]$ are extracted, following the approach for fixation type 1 as described in Eq. 5 and 6 for $c_k = 1$ in Section IV-A4. For fixation type 2, features, $f_i^{c=2}$ and $g_i^{c=2}$, of pixel p_i located within the distance of ρ_1 from L_k are extracted using Eq. 5 and 6 for $c_k = 2$. However, extracting features for every p_i can be very slow. We therefore employ the pixels at the intersections of L_k, plus a further 100 random points from a set of p_i.

These features are input to the SVM using both models to predict the probability P of corresponding to an eye position. The probability map S_1 of type 1 is generated by combining the probability $P_r^{c=1}$ from all N_r regions, i.e. $S_1 = \sum_{r=1}^{N_r} B_r P_r^{c=1}$, where B_r is a binary mask for region r. For type 2, the probabilities of the rest of p_i are interpolated from those selected points for feature extraction. Finally, S_1 and S_2 are combined as described in Eq. 8. The weight α is calculated using terrain difficulty estimation.

$$S_k = \begin{cases} S_1, & \text{if } S_2 = 0 \\ S_2, & \text{if } S_1 = 0 \\ \alpha \cdot S_1 + (1-\alpha) \cdot S_2, & \text{otherwise} \end{cases} \quad (8)$$

As discussed in Section III, when the difficulty of the terrain increases, humans fixate more often at object boundaries. We score the terrain difficulty using the high spatial frequencies in the current frame and how they are distributed throughout the frame. The weight α is computed using Eq. 9.

$$\alpha' = \text{var}(\Psi_k) \cdot \text{mean}(\Psi_k) \quad (9a)$$

$$\alpha = \frac{1}{1 + e^{-k(\alpha' - \mu_\alpha)}} \quad (9b)$$

where Ψ_k is the mean of high-pass magnitudes of frame k, mean(x) and var(x) are mean and variance of data x. The large mean of high-pass magnitudes implies the presence of strong structures, edges or corners. The large variance implies that the materials may be rough. μ_α is the mean of all α' which is 1.05×10^{-3}. k is the steepness of the curve. We set k to 2.70×10^3, where $\alpha = 0.1$ and 0.9 at $\frac{\min(\alpha')}{2}$ and $\max(\alpha') + \frac{\min(\alpha')}{2}$, respectively. These values are computed from all sequences. Finally, the low-pass Gaussian filter is applied in order to smooth the result map.

TABLE II
LIST OF PARAMETERS USED IN OUR METHOD

parameter	symbol	value
frame height	H	960 pixels
frame width	W	1280 pixels
number of backward/forward frames	N_b/N_f	15 frames
sharpness weight	α	0.9
distance from region boundary	d_{L_k}	10 pixels
radius of area around a eye position	ρ_1	20 pixels
outer radius of surrounding area	ρ_2	60 pixels

V. Results and discussion

We tested the proposed scheme using 8 sequences with eye tracking from 8 participants. Each sequence contains four terrain types (*flat concrete*, *slant cobbles*, *stepping stones* and *rocks*) as shown in Fig. 1, and they vary between approximately 4-6 minutes in duration.

A. Dominant features and classification performance

We first studied which features were dominant in human perception for locomotion using a sequential forward selection (SFS) [51] and a normal-based feature selection (NFS) [52]. The experimental results of the SFS and NSF tests show that kurtosis and entropy are the best features of the intensity level distributions for all terrain types and for both fixation types. Kurtosis measures the peakedness of the distribution and the heaviness of its tail [53], while entropy measures the randomness of the texture. For the wavelet features, the magnitudes of combined subbands of the decomposition level 3 are dominant for all terrain types and for both fixation types. This could be because it is the best level for capturing structure of the image, which agrees with our findings in previous work [54]. However, comparing amongst wavelet orientations, no distinctive features were present for all terrain types. The features of horizontal and diagonal directions are dominant in the stepping stone sequences, while the features of vertical directions are more prominent in the rock sequences. For LBP, the bins of the histogram that indicate rougher texture are used more in the *slant cobbles* and *rocks* sequences. It is obvious that the eye positions are highly task-dependent, which agrees with the study by [7]. Easy terrain contains relatively little salient visual information compared to uneven surfaces, particularly the *stepping stones* and *rocks*, where the characteristics may be more important for maintaining posture balance.

Table III compares the classification accuracies using the actual texture features (f_k), the difference between those of eye position and its surroundings (g_k), and both (f_k,g_k). g_k gives significantly better results when classifying eye positions

TABLE III
CLASSIFICATION ACCURACY (%) USING ONE EYE POSITION PER FRAME

terrain	f_k	g_k	(f_k, g_k)
Flat concrete	65.21	85.43	88.47
Slant cobbles	64.50	91.28	93.56
Stepping stones	70.65	64.62	78.82
Rocks	97.68	90.49	92.93
all types	74.51	82.96	88.44

and random points than f_k for *flat concrete*, *slant cobbles* and *rocks* with up to 25% improvement in classification accuracy, while f_k gives better results (by 6%) for the case of *stepping stones*. Using all features give the best classification results with no significant increase in computational time. Moreover, using both f_k and g_k can differentiate between fixation types with 99.9% accuracy for all terrains, while using f_k alone can achieve only 77%. Therefore, we suggest using all features (f_k, g_k).

B. Priority map

The objective results are evaluated using a receiver operating characteristic (ROC curve). A ground truth for each frame is constructed using N_f forward and N_b backward frames in order to allow some misalignments in temporal direction. We define a score $s_{k,m}$ for each warped point $p_{k,m}, m \in [k-N_b, k+N_f], m \neq 0$, according to the distance between frames ($d_m = k - m$) as the eye positions in the frames further from the current one had lower probabilities of being an eye position in the current frame. We define the score of the eye position of the current frame to 1 (maximum) and the scores of other frames are linearly decayed until that of the furthest frame which is set to equal to 0.5. That is, $s_{k,m}$ given to sample $p_{k,m}$ is equal to $1 - \frac{d_m}{N_b+N_f}$. Then, a 2-dimensional Gaussian function ($\sigma = 10$) is implemented as a point spread function to allow some spatial shifts of positions from the ground truth. This is similar to constructing a psychophysical fixation map, where a Gaussian-distributed activity is assumed [17].

In addition, other objective results are computed using i) normalized scanpath saliency (NSS), ii) linear correlation (CC), iii) Earth Mover's Distance (EMD), iv) histogram intersection or similarity (SIM), where EMD=0 for identical distribution, v) Area Under ROC curve measure based on Ali Borji's method (AUC-Borji) [55], and vi) Area Under ROC curve measure based on Judd's method (AUC-Judd) [56].

1) Proposed model testing: We first investigated the performance of the proposed method for individual participants. A 2-fold cross validation was employed - the first half of the sequence was used for training and the second half was used for evaluation. Then, they were swapped and the results were averaged. Table IV shows the average of the areas under the ROC curve computed by Ali Borji's method and Judd's method. The last column and the last row shows the means and the standard deviations of the results for each terrain type and each participant, respectively. The proposed method performs the best for Participant 2 (highest mean AUC and lowest standard deviation), while it performs the worst for Participant 8 (lowest mean AUC and high standard deviation).

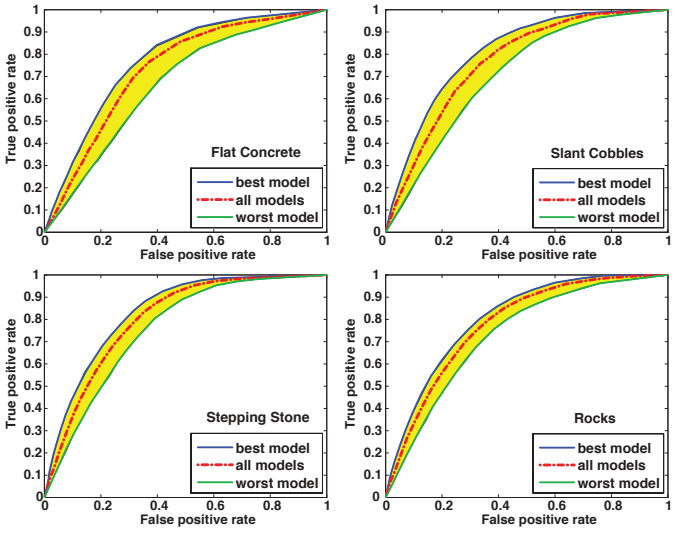

Fig. 8. ROC curves of the proposed method when using one sequence for training.

Overall, the proposed method achieves consistent performance for individual participants as the standard deviation is not high.

We tested whether our proposed model could use the data from one participant to predict the eye positions of the others. We therefore trained the classifier using one sequence, and tested with the others. Fig. 8 shows the ROC curves that plot the results from the best model, the worst model and the average from all models. The yellow areas (representing the gap between the worst and the best models) demonstrate the variations of the results when different models are employed. A small (narrow) area implies that the proposed approach is generalized and should give similar results when applied to different data. From Fig. 8, the yellow area of the *flat concrete* plot is the largest because participants tend to look around more often, but the resultant eye positions cannot be detected as outliers because they are located within the central 80% of the population distribution. Such data are not suitable for training because of large amount of noise. In contrast, the result of the rocky terrain shows the least variation. This is because every participant had to concentrate on their path, so the fixation patterns are relatively similar.

The areas above the average lines are smaller than those below for all terrains. This implies that the proposed method has commonality with most sequences and that there are some outlier sequences which result in lower performance. These sequences should be excluded from the training process for real-world use.

2) Performance comparison: Here we used three sequences for testing and the rest for training the classifiers, so there were in total 56 cross-validation tests. We compared our results to those of i) proto-objects [57], ii) DT-CWT [58], iii) image signature [18], iv) region covariances (CovSal) [59], v) multi-level features (Judd) [25], vi) spectral saliency detector (SSD) [23], vii) learning discriminative subspaces (LDS) [28], and viii) Deep convolutional network (SalNet) [29]. These methods are in the top rank of the MIT saliency benchmark that provide accessible code [30]. The saliency proto-objects are extracted

TABLE IV
PERFORMANCE OF THE PROPOSED METHOD ON THE INDIVIDUAL PARTICIPANT MEASURED BY THE AVERAGE BETWEEN AUC-BORJI AND AUC-JUDD

Terrain	#1	#2	#3	#4	#5	#6	#7	#8	mean±std
Concrete	0.72	0.85	0.81	0.82	0.88	0.76	0.71	0.73	0.79±0.065
Cobbles	0.87	0.88	0.88	0.90	0.82	0.79	0.81	0.79	0.84±0.046
Stones	0.88	0.89	0.88	0.90	0.90	0.84	0.90	0.87	0.88±0.021
Rocks	0.84	0.88	0.85	0.88	0.86	0.83	0.89	0.86	0.86±0.022
mean±std	0.83±0.072	0.88±0.021	0.86±0.029	0.87±0.039	0.86±0.033	0.81±0.037	0.83±0.088	0.81±0.067	0.84±0.054

using rarity and contrast within the segmented region. The DT-CWT method captures visual priority from the energy in the wavelet domain, whilst the image signature method and the SSD method employ the DCT and the DFT, respectively. The CovSal method computes covariance between non-overlapping patches. The Judd and SalNet methods are based on machine learning and were trained with our dataset. Fig 9 shows the ROC curves of each terrain and Table V shows the objective results. For fair comparison, we also show the results of our method when the centre-bias approach is included. We simply applied a Gaussian weight with σ=100.

The proto-objects and image signature methods do not perform well for any terrain. These two methods were originally intended for object detection purposes. Hence, when there are no distinctive objects in the walking scene, they pick up the areas with most salient information, often not related to locomotion. The DT-CWT method performed slightly better, but it cannot be employed for a sequence with low energy at high frequencies, such as the flat terrain. The performance of the CovSal, Judd, SSD and LDS methods are close to ours. This is mainly because these methods apply a centre bias, which is usually justified for most eye fixation experiments since humans do not simply rely on eye movements, but also head movements to improve vision. However, this may not be practical for autonomous machines or robots if their visual inputs cannot be moved automatically, or there is insufficient knowledge of the scene to enable safe traversal. For a fairer comparison, when we integrated a center bias into our method, it outperformed the CovSal, Judd, SSD and LDS methods by approximately 8%, 12%, 8% and 6% (3%, 8%, 3% and 2% without centre bias), respectively, computed from the average of all metrics as shown in Table V, excluding EMD. These results clearly show that our method can be applied to machines both with and without mechanisms for controlling head or camera movement. SalNet shows comparable performance to our method without centre bias (but 4% less if our method applied centre bias). This could be because the deep convolutional network works in the way that replicates the primate visual system. The model was created by self-learning from visual information participants perceived. However, this method requires much more training data to ensure generalisation.

The estimated priority maps are shown in Fig. 10. The fixations for *flat concrete* are obviously difficult to predict. The eye positions on this simple terrain are generally in the distance, sometimes on the walking path and sometimes on the surroundings. The priority maps give similar results at the far distance, but not for the near areas. For more complex terrains, where humans concentrate on searching for a safe traversal, our results achieve the best estimation. The CovSal method produces obvious centre-bias results, while the Judd method gives more spread out probability maps than others.

3) Exploitation of temporal relations: In this section, we exploited temporal information by warping eye positions of neighbouring frames onto the current frame. This accumulates key points from previous frames and also those points that would appear in the short-term or immediate future. These warping locations, including the current one, are labelled as positive against the random points labelled as negative, and they are used to train the classifiers. In addition, we also tested by warping the eye position of the current frames to the neighbouring frames. The features of the corresponding points are combined with weighted average - smaller weights are applied to further frames. This approach should decrease the effect of varying orientations due to walking [45]. For both cases, we tested using 5 and 20 neighbouring frames. However, the results were not significantly different to those of our original approach using only the eye position of the current frame, giving only slight improvement on easy terrains, i.e. *flat concrete* and *slant cobbles*, and no improvement on complex terrains, i.e. *stepping stones* and *rocks*. This is because the fixations occur instantly according to the incoming terrain during locomotion. However, there is evidence that long-duration fixations and revisited fixations used for path planning can improve the performance of instant fixation prediction [60].

A long short-term memory (LSTM) system could potentially improve overall system performance as it is suitable for tasks where there are time lags of unknown size and bound between important events [34]. Applying LSTM (or any RNN) to our application is however complicated, because the scenes in our videos are perspective with background constantly changing and affected from gait bounce signals due to the body vaults over the leg. In general, visual information from previous frames should be projected to the current frame geometry so as to have similar characteristics. This means that the signals in the hidden states of an RNN may need to be non-linearly transformed. Moreover, previously LSTMs have been designed for applications where predictions require knowledge of what happened in the past, such as language processing where a prediction of the next word in a sentence can be done with the knowledge of what the previous words are. In contrast, achieving safe locomotion using visual information may rely on upcoming event more than on the past. Therefore, to apply LSTM to our application, an intensive study is required which will be the subject of future research.

4) Robustness: We applied the model producing the best result in Section V-B2 to cases of more difficult terrains. Twelve sequences with eye tracking data were employed. They were captured from twelve participants walking in two

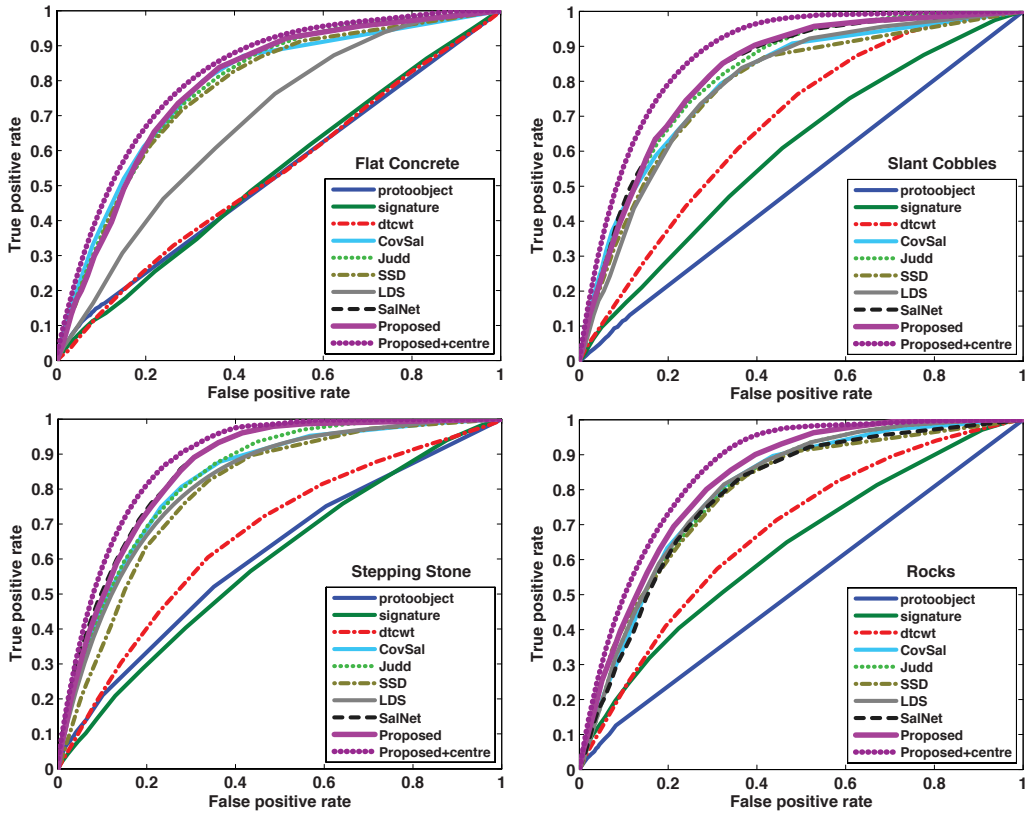

Fig. 9. Performance comparison using ROC curve for (Top-left) flat concrete, (Top-right) slant cobbles, (Bottom-left) stepping stones, and (Bottom-right) rocks.

TABLE V
PERFORMANCE COMPARISON USING SEVERAL METRICS FOR SALIENCY ASSESSMENT

terrain	method	Proto-object	Signature	DTCWT	CovSal	Judd	SSD	LDS	SalNet	Proposed	Proposed+centre
Flat concrete	NSS	0.16	0.15	0.13	1.21	1.12	**1.23**	0.80	1.22	1.19	**1.29**
	CC	0.07	0.06	0.04	**0.45**	0.38	**0.51**	0.32	0.40	0.35	0.38
	EMD	10.25	12.41	12.41	**7.76**	10.93	8.30	8.22	8.25	8.50	**8.21**
	SIM	0.27	0.14	0.13	**0.54**	0.19	0.51	0.42	**0.62**	0.40	0.51
	AUD-Borji	0.52	0.52	0.53	0.70	**0.79**	0.72	0.67	0.78	**0.79**	**0.81**
	AUD-Judd	0.60	0.58	0.57	**0.83**	0.82	0.76	0.76	**0.83**	**0.83**	**0.85**
Slant cobbles	NSS	0.02	0.48	0.66	**1.31**	1.21	1.25	1.26	1.29	**1.31**	**1.64**
	CC	0.01	0.13	0.19	0.45	0.37	**0.53**	**0.50**	0.48	0.41	**0.50**
	EMD	11.40	12.36	12.34	8.26	11.37	**7.83**	8.13	**6.96**	8.10	8.06
	SIM	0.13	0.13	0.13	**0.43**	0.16	0.32	0.33	0.33	**0.39**	0.32
	AUD-Borji	0.50	0.62	0.69	0.74	0.81	0.72	0.77	0.82	**0.83**	**0.89**
	AUD-Judd	0.58	0.70	0.74	0.84	0.85	0.78	0.83	**0.89**	0.86	**0.91**
Stepping stones	NSS	0.37	0.39	0.52	1.39	1.32	1.20	1.54	**1.98**	1.78	**2.11**
	CC	0.11	0.10	0.15	0.41	0.36	0.49	**0.63**	0.49	0.45	**0.54**
	EMD	9.58	12.39	12.65	7.47	11.21	7.52	6.29	6.57	**5.48**	**5.48**
	SIM	0.32	0.11	0.10	0.41	0.14	**0.52**	0.51	**0.52**	**0.52**	**0.52**
	AUD-Borji	0.53	0.61	0.65	0.73	0.83	0.73	0.81	**0.87**	0.86	**0.88**
	AUD-Judd	0.65	0.71	0.72	0.86	0.87	0.80	0.86	**0.90**	**0.90**	**0.92**
Rocks	NSS	0.10	0.68	0.61	1.01	1.10	1.25	**1.40**	1.06	1.25	**1.46**
	CC	0.02	0.20	0.17	0.31	0.30	0.40	**0.54**	0.35	0.40	**0.45**
	EMD	11.47	12.32	13.05	**8.44**	11.86	8.70	8.74	10.32	8.71	**8.59**
	SIM	0.13	0.13	0.10	0.31	0.13	0.35	0.33	0.26	**0.36**	**0.37**
	AUD-Borji	0.51	0.66	0.67	0.72	0.80	0.74	0.80	0.76	**0.85**	**0.86**
	AUD-Judd	0.61	0.74	0.73	0.84	0.84	0.82	0.85	0.82	**0.87**	**0.88**
All	NSS	0.16	0.43	0.48	1.23	1.19	1.23	1.25	**1.39**	1.38	**1.63**
	CC	0.05	0.13	0.14	0.42	0.35	**0.48**	**0.50**	0.43	0.40	**0.48**
	EMD	10.67	12.37	12.61	7.98	11.34	8.09	7.85	8.02	**7.70**	**7.58**
	SIM	0.21	0.13	0.12	0.41	0.16	**0.43**	0.40	**0.42**	0.41	**0.42**
	AUD-Borji	0.52	0.60	0.64	0.72	0.81	0.73	0.76	0.81	**0.83**	**0.86**
	AUD-Judd	0.61	0.68	0.69	0.84	0.84	0.79	0.83	**0.86**	**0.86**	**0.89**

environments, six in the 'woods' and six in the 'park', as shown in Fig. 11. The woods sequences contained various

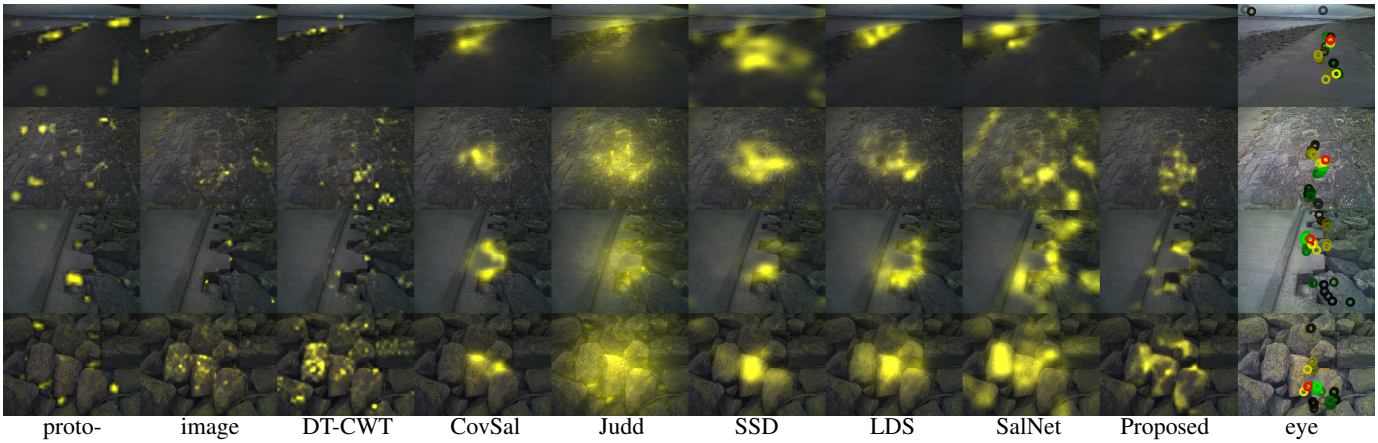

Fig. 10. Priority maps generated using i) proto-objects, ii) image signature, iii) DT-CWT, iv) CovSal, v) Judd, vi) SSD, vii) LDS, viii) SalNet, and ix) our proposed method. Right images show eye position at the current frame (red) and warped eye positions from the backward frames (green) and the forward frames (yellow) - darker colours represent fixations from frames further from the current frame.

TABLE VI
SALIENCY ASSESSMENT WHEN APPLYING TO WOODS AND PARK TERRAINS (AVERAGE VALUES)

method	NSS	CC	EMD	SIM	AUD-Borji	AUD-Judd
Proto-object	0.12	0.08	14.17	0.25	0.51	0.56
Signature	0.35	0.11	14.44	0.13	0.56	0.77
DTCWT	0.33	0.13	12.41	0.13	0.63	0.68
CovSal	1.35	0.39	10.33	**0.44**	0.77	0.89
Judd	1.22	0.32	12.65	0.15	**0.81**	0.90
SSD	1.08	0.11	11.22	0.27	0.72	0.86
LDS	1.43	0.14	9.86	0.32	0.79	0.89
SalNet	0.82	0.30	12.94	0.12	0.68	0.79
Proposed	**1.47**	**0.42**	**8.52**	**0.44**	0.80	**0.91**
Proposed+ center	**1.63**	**0.45**	**7.94**	**0.45**	**0.85**	**0.92**

Fig. 11. Eye tracking woods (top row) and park (bottom row) sequences containing a variety of ground materials. The circles show fixated points.

sloped terrains and a mix of materials including dirt, rocks, grass, and woods. The park sequences also contained a variety of materials, but the walking paths were flatter and more winding. These sequences vary between approximately 4-6 minutes in duration. Fig. 12 shows the estimated priority maps overlaid on the images. In the case of complex terrain, almost everywhere in the scene has high saliency leading to difficulty in prioritization. We can see that the results of the image signature, DT-CWT and Judd show bright yellow areas all over the images. Our priority maps show the most relevant areas to the ground truths.

Table VI shows the numerical results with our method achieving the best values. This confirms the robustness of our approach and our model is not overfitted, as the model was trained on the different terrain types. In contrast, the performance of SalNet is significantly reduced. This might be due to overfitting as deep neural nets require much more training data to perform well. With the good performance of our proposed method on this scenario, it can also confirm that the observation of two-type fixations can also be found in different terrains.

VI. CONCLUSIONS AND FUTURE WORK

We have presented a novel priority estimation method for human fixations during locomotion. Sequences with eye trackers captured while participants traversed four different terrains were analysed using texture features. We create priority maps from two types of fixations. The first relates to the eye positions of the safe places to step. The second is where the fixations are located near or on the boundary lines of the segmentation map. This indicates where participants are aware of borders and terrain changes. The local texture features at the eye positions are employed to train an SVM classifier. Our proposed approach outperforms existing methods for complex terrains and gives similar results on simple terrains. However, unlike existing methods, we do not apply a centre bias assumption, so our approach should perform better when using a fixed-position camera. Additionally, our results show significant improvement over existing methods when a centre bias constraint is applied, providing flexibility for use in

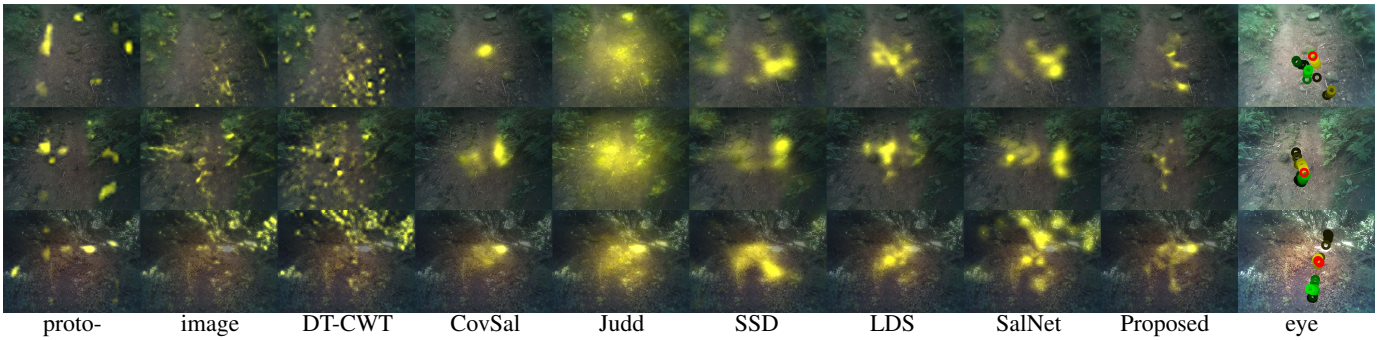

Fig. 12. Priority maps of woods (row 1 and 2) and park (row 3) terrains generated using i) proto-objects, ii) image signature, iii) DT-CWT, iv) CovSal, v) Judd, vi) SSD, vii) LDS, viii) SalNet, and ix) our proposed method. Right images show eye position at the current frame (red) and warped eye positions from the backward frames (green) and the forward frames (yellow) - darker colours represent fixations from frames further from the current frame.

applications where the camera can be rapidly repositioned to improve vision.

For future work, the enhanced system will be validated by comparing fixation points and features across humans and machines for a range of scenarios. Subsequently, a hierarchical classifier based on levels of terrain complexity associated with a recurrent neural network will be developed to achieve both high accuracy across a diverse set of terrains and faster computational performance.

REFERENCES

[1] M. Hayhoe and D. Ballard, "Eye movements in natural behavior," *TRENDS in Cognitive Sciences*, vol. 9, pp. 188–194, 2005.

[2] M. F. Land, "Eye movements and the control of actions in everyday life," *Progress in Retinal and Eye Research*, vol. 25, no. 3, pp. 296–324, 2006.

[3] A. E. Patla and J. N. Vickers, "How far ahead do we look when required to step on specific locations in the travel path during locomotion?" *Experimental Brain Research*, vol. 148, no. 1, pp. 133–138, 2003.

[4] S. Fotios, J. Uttley, C. Cheal, and N. Hara, "Using eye-tracking to identify pedestrians critical visual tasks, part 1. dual task approach," *Lighting research & technology*, vol. 47, pp. 133–148, 2015.

[5] A. E. Patla and J. N. Vickers, "Where and when do we look as we approach and step over an obstacle in the travel path?" *Neurophysiology*, vol. 8, no. 17, pp. 3661–3665, 1997.

[6] J. M. Franchak and K. E. Adolph, "Visually guided navigation: Head-mounted eye-tracking of natural locomotion in children and adults," *Vision Research*, vol. 50, pp. 2766–2774, 2010.

[7] D. Marigold and A. Patla, "Gaze fixation patterns for negotiating complex ground terrain," *Neuroscience*, vol. 144, pp. 302–313, 2007.

[8] M. A. Hollands, A. E. Patla, and J. N. Vickers, "look where youre going!: gaze behaviour associated with maintaining and changing the direction of locomotion," *Experimental Brain Research*, vol. 143, no. 2, pp. 221–230, 2002.

[9] K. Kitazawa and T. Fujiyama, "Pedestrian vision and collision avoidance behavior: Investigation of the information process space of pedestrians using an eye tracker," *Pedestrian and Evacuation Dynamics 2008*, pp. 95–108, 2009.

[10] J. J. Gibson, "Visually controlled locomotion and visual orientation in animals," *British Journal of Psychology*, vol. 49, no. 3, pp. 182–194, 1958.

[11] J. H. Fecteau and D. Munoz, "Salience, relevance, and firing: a priority map for target selection," *TREN*, vol. 10, no. 8, pp. 382–390, 2006.

[12] G. J. Zelinsky and J. W. Bisley, "The what, where, and why of priority maps and their interactions with visual working memory," *Annals of the New York Academy of Sciences*, vol. 1339, pp. 154–164, 2015.

[13] C. Chang and C. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011.

[14] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," *Human Neurobiology*, vol. 4, no. 4, pp. 219–227, 1985.

[15] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[16] D. Walther and C. Koch, "Modeling attention to salient proto-objects," *Neural Networks*, vol. 19, no. 9, pp. 1395–1407, 2006.

[17] Q. Zhao and C. Koch, "Learning a saliency map using fixated locations in natural scenes," *Journal of Vision*, vol. 11, no. 3, pp. 1–15, 2011.

[18] X. Hou, J. Harel, and C. Koch, "Image signature: Highlighting sparse salient regions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 1, pp. 194–201, 2012.

[19] A. Borji, D. Sihite, and L. Itti, "What/where to look next? modeling top-down visual attention in complex interactive environments," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 44, no. 5, pp. 523–538, 2014.

[20] N. Murray, M. Vanrell, X. Otazu, and C. Parraga, "Saliency estimation using a non-parametric low-level vision model," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 433–440.

[21] L. Zhang, M. Tong, T. Marks, H. Shan, and G. Cottrell, "Sun: A bayesian framework for saliency using natural statistics," *Journal of Vision*, vol. 8, pp. 1–20, 2008.

[22] M. Jian, K. M. Lam, J. Dong, and L. Shen, "Visual-patch-attention-aware saliency detection," *IEEE Transactions on Cybernetics*, vol. 45, no. 8, pp. 1575–1586, Aug 2015.

[23] J. Li, L. Y. Duan, X. Chen, T. Huang, and Y. Tian, "Finding the secret of image saliency in the frequency domain," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 12, pp. 2428–2440, Dec 2015.

[24] N. D. B. Bruce and J. K. Tsotsos, "Saliency, attention, and visual search: An information theoretic approach," *Journal of Vision*, vol. 9, no. 3, pp. 1–24, 2009.

[25] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *IEEE International Conference on Computer Vision*, 2009, pp. 2106–2113.

[26] Q. Wang, Y. Yuan, P. Yan, and X. Li, "Saliency detection by multiple-instance learning," *IEEE Transactions on Cybernetics*, vol. 43, no. 2, pp. 660–672, April 2013.

[27] M. Liang and X. Hu, "Feature selection in supervised saliency prediction," *IEEE Transactions on Cybernetics*, vol. 45, no. 5, pp. 914–926, May 2015.

[28] S. Fang, J. Li, Y. Tian, T. Huang, and X. Chen, "Learning discriminative subspaces on random contrasts for image saliency analysis," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–14, 2016.

[29] J. Pan, E. Sayrol, X. Giro-I-Nieto, K. McGuinness, and N. E. OConnor, "Shallow and deep convolutional networks for saliency prediction," in *IEEE Conference on Computer Vision and Pattern Recognition*, June 2016, pp. 598–606.

[30] Z. Bylinskii, T. Judd, A. Borji, L. Itti, F. Durand, A. Oliva, and A. Torralba, "Mit saliency benchmark."

[31] L. Itti, N. Dhavale, and F. Pighin, "Realistic avatar eye and head animation using a neurobiological model of visual attention," *Proceedings of SPIE International Symposium on Optical Science and Technology*, vol. 5200, pp. 64–78, 2003.

[32] O. Le Meur, D. Thoreau, P. Le Callet, and D. Barba, "A spatio-temporal model of the selective human visual attention," in *IEEE International Conference on Image Processing*, vol. 3, 2005, pp. 1188–1191.

[33] O. Boiman and M. Irani, "Detecting irregularities in images and in video," *International Journal of Computer Vision*, vol. 74, no. 1, pp. 17–31, 2005.

[34] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[35] x. wang, l. gao, J. Song, and H. Shen, "Beyond frame-level cnn: Saliency-aware 3d cnn with lstm for video action recognition," *IEEE Signal Processing Letters*, no. 99, pp. 1–5, 2016.

[36] P. Rodriguez, G. Cucurull, J. Gonzlez, J. M. Gonfaus, K. Nasrollahi, T. B. Moeslund, and F. X. Roca, "Deep pain: Exploiting long short-term memory networks for facial expression classification," *IEEE Transactions on Cybernetics*, pp. 1–11, 2017.

[37] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, pp. 185–207, 2013.

[38] E. Awh, A. V. Belopolsky, and J. Theeuwes, "Top-down versus bottom-up attentional control: a failed theoretical dichotomy," *Trends in Cognitive Sciences*, vol. 16, no. 8, pp. 437–443, 2012.

[39] B. Hart and Einhäuser, "Mind the step: complementary effects of an implicit task on eye and head movements in real-life gaze allocation," *Experimental Brain Research*, vol. 223, pp. 233–249, 2012.

[40] S. Kastner, P. D. Weerd, and L. G. Ungerleider, "Texture segregation in the human visual cortex: A functional MRI study," *Journal of Neurophysiology*, vol. 83, no. 4, pp. 2453–2457, 2000.

[41] T. Foulsham, E. Walker, and A. Kingstone, "The where, what and when of gaze allocation in the lab and the natural environment," *Vision Research*, vol. 51, pp. 1920–1931, 2011.

[42] N. Anantrasirichai, I. D. Gilchrist, and D. R. Bull, "Fixation identification for low-sample-rate mobile eye trackers," in *IEEE International Conference on Image Processing*, 2016, pp. 3126–3130.

[43] P. Hill, N. Anantrasirichai, A. Achim, M. Al-Mualla, and D. Bull, "Undecimated dual tree complex wavelet transforms," *Signal Processing: Image Communication*, vol. 35, pp. 61–70, 2015.

[44] R. O'Callaghan and D. Bull, "Combined morphological-spectral unsupervised image segmentation," *IEEE Transactions on Image Processing*, vol. 14, no. 1, pp. 49–62, 2005.

[45] N. Anantrasirichai, J. Burn, and D. Bull, "Terrain classification from body-mounted cameras during human locomotion," *IEEE Transactions on Cybernetics*, vol. 45, no. 10, pp. 2249–2260, 2015.

[46] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971 –987, 2002.

[47] J. Jones and L. Palmer, "An evaluation of the two-dimensional gabor filter model of simple receptive fields in cat striate cortex," *Journal of Neurophysiology*, vol. 58, no. 6, pp. 1233–1258, 1987.

[48] D. Gao, V. Mahadevan, and N. Vasconcelos, "On the plausibility of the discriminant centersurround hypothesis for visual saliency," *Journal of Vision*, vol. 8, no. 7, pp. 1–18, 2008.

[49] B. W. Tatler, R. J. Baddeley, and I. D. Gilchrist, "Visual correlates of fixation selection: Effects of scale and time," *Vision Research*, vol. 45, pp. 643–659, 2005.

[50] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2004.

[51] D. W. Aha and R. L. Bankert, "A comparative evaluation of sequential feature selection algorithms," *Learning from Data: Lecture Notes in Statistics*, vol. 112, pp. 199–206, 1996.

[52] J. Brank, M. Grobelnik, N. Milić-frayling, and D. Mladenić, "Feature selection using support vector machines," in *Proc. of the International Conference on Data Mining Methods and Databases for Engineering, Finance, and Other Fields*, 2002, pp. 84–89.

[53] K. P. Balanda and H. L. MacGillivray, "Kurtosis: A critical review," *The American Statistician*, vol. 42, no. 2, pp. 111–119, 1988.

[54] N. Anantrasirichai, J. Burn, and D. Bull, "Orientation estimation for planar textured surfaces based on complex wavelets," in *IEEE International Conference on Image Processing*, 2014, pp. 3372–3376.

[55] A. Borji, H. Tavakoli, D. Sihite, and L. Itti, "Analysis of scores, datasets, and models in visual saliency prediction," in *IEEE International Conference on Computer Vision*, 2013, pp. 921–928.

[56] N. Riche, M. Duvinage, M. Mancas, B. Gosselin, and T. Dutoit, "Saliency and human fixations: State-of-the-art and study of comparison metrics," in *IEEE International Conference on Computer Vision*, 2013, pp. 1153–1160.

[57] V. Yanulevskaya, J. Uijlings, J.-M. Geusebroek, N. Sebe, and A. Smeulders, "A proto-object-based computational model for visual saliency," *Journal of Vision*, vol. 13, no. 13, pp. 1–19, 2013.

[58] J. Fauqueur, N. Kingsbury, and R. Anderson, "Multiscale keypoint detection using the dual-tree complex wavelet transform," in *IEEE International Conference on Image Processing*, Oct. 2006, pp. 1625–1628.

[59] E. Erdem and A. Erdem, "Visual saliency estimation by nonlinearly integrating features using region covariances," *Journal of Vision*, vol. 13, no. 4, pp. 1–20, 2013.

[60] N. Anantrasirichai, I. D. Gilchrist, and D. R. Bull, "Visual salience and priority estimation for locomotion using a deep convolutional neural network," in *IEEE International Conference on Image Processing*, 2016, pp. 1599–1603.